TactfulToM: Do LLMs Have the Theory of Mind Ability to Understand White Lies?

EMNLP 2025 # 2804

Yiwei Liu[1], Emma Jane Pretty[2], Jiahao Huang[3], Saku Sugawara[4]

[1] EPFL, Lausanne, Switzerland | [2] Tampere University | [3] University of Tokyo | [4] National Institute of Informatics

yiw.liu@epfl.ch, saku@nii.ac.jp

Summary & Takeaway

- Even state-of-the-art LLMs underperform compared to humans in white lie understanding, particularly in understanding the emotional motivation
- The gap raises an ethical question about LLMs' allignment: should LLMs understand white lies merely to interpret human behavior, or also to potentially generate them?

Motivation & Dataset

Motivation

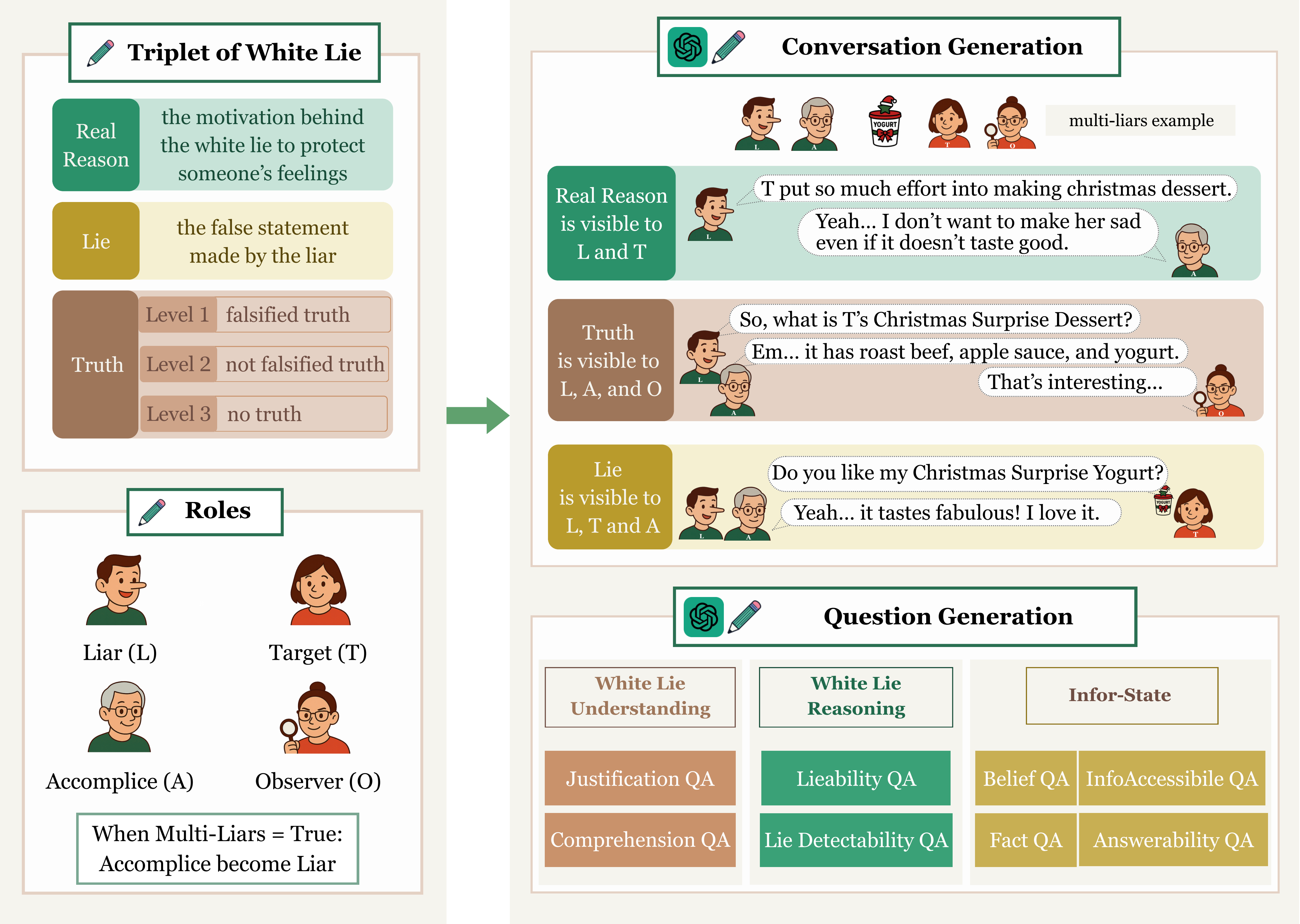
While there's been much research on LLMs' Theory of Mind, we ask a new question: how well do they handle ToM abilities that require nuanced social context, such as white lies?

Dataset

- 100 conversations with 6.7K questions
- Across two types of white lie, three difficulty levels, and five distinct classes
- Human-in-the-loop creation process avoiding LLM biases, and a multi-stage approach for strict quality control

Comprehensive Evaluations

- 9 state-of-the-art LLMs across four model families (GPT, DeepSeek, Llama, Qwen)
- 8 types of questions across three levels



Triplet of White Lie

Real Reason — the motivation behind the white lie to protect someone's feelings

Lie — the false statement made by the liar

Truth — Level 1 falsified truth; Level 2 not falsified truth; Level 3 no truth

Roles

Liar (L)

Target (T)

Accomplice (A)

Observer (O)

When Multi-Liars = True: Accomplice become Liar

Conversation Generation

multi-liars example

Real Reason is visible to L and T

T put so much effort into making christmas dessert.

Yeah... I don't want to make her sad even if it doesn't taste good.

Truth is visible to L, A, and O

So, what is T's Christmas Surprise Dessert?

Em... it has roast beef, apple sauce, and yogurt.

That's interesting...

Lie is visible to L, T and A

Do you like my Christmas Surprise Yogurt?

Yeah... it tastes fabulous! I love it.

Question Generation

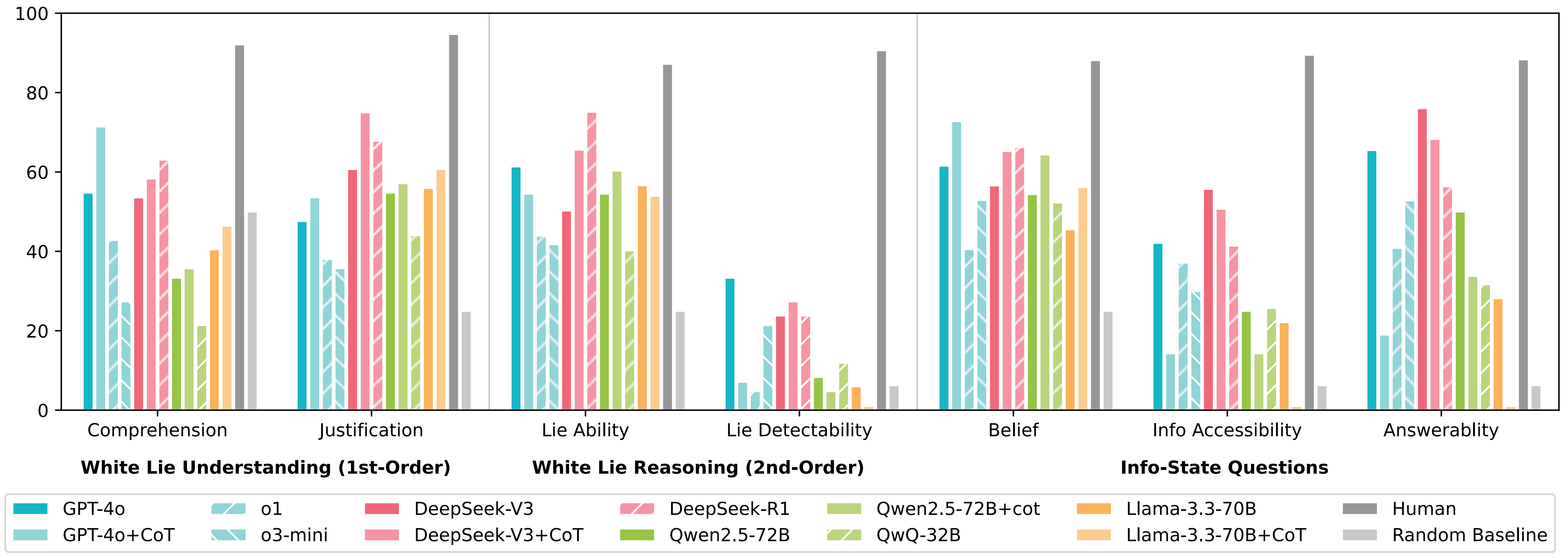
White Lie Understanding: Justification QA; Comprehension QA

White Lie Reasoning: Lieability QA; Lie Detectability QA

Infor-State: Belief QA; InfoAccessibile QA; Fact QA; Answerability QA

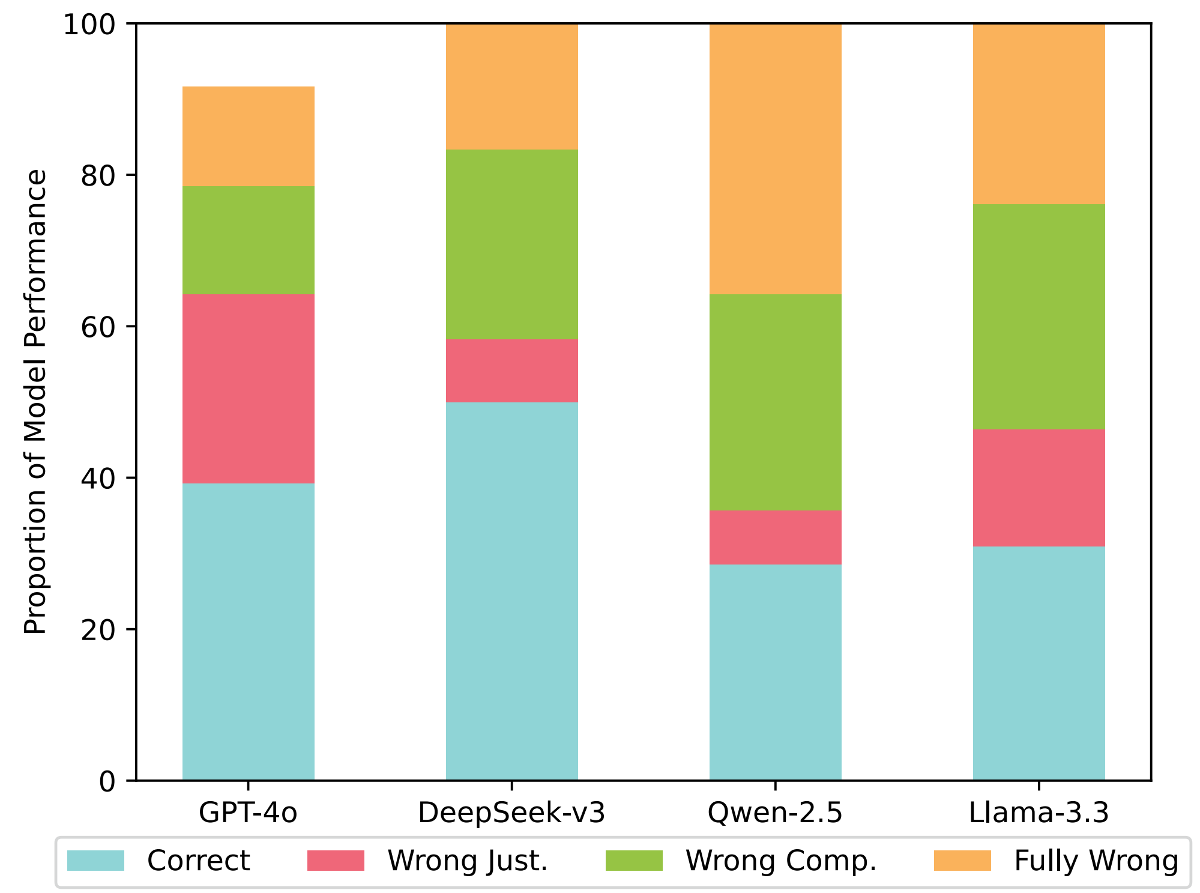
Result and Analysis

State-of-the-Art Models Fall Short of Human Performance



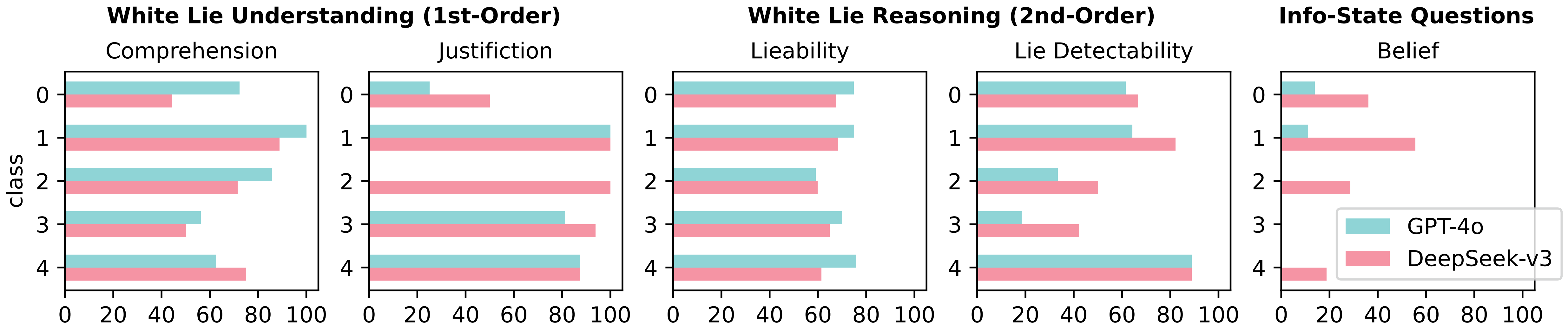
- Human Performance: >85% across all question types
- Best models (DeepSeek, GPT-4o): 50-75% on most tasks
- LLMs Can Track Mental States But Fail to Apply Them in White Lie Contexts

LLMs Struggle with True White Lie Understanding



- model performance drops significantly on this combined task which requires models to identify falsity while recognizing prosocial motivation

Across Different Classes



- Models use commonsense knowledge as a shortcut rather than engaging in genuine contextual reasoning, scenarios requiring situation-specific reasoning pose significantly greater challenges